Update on International HPC Activities

(mostly Asia)

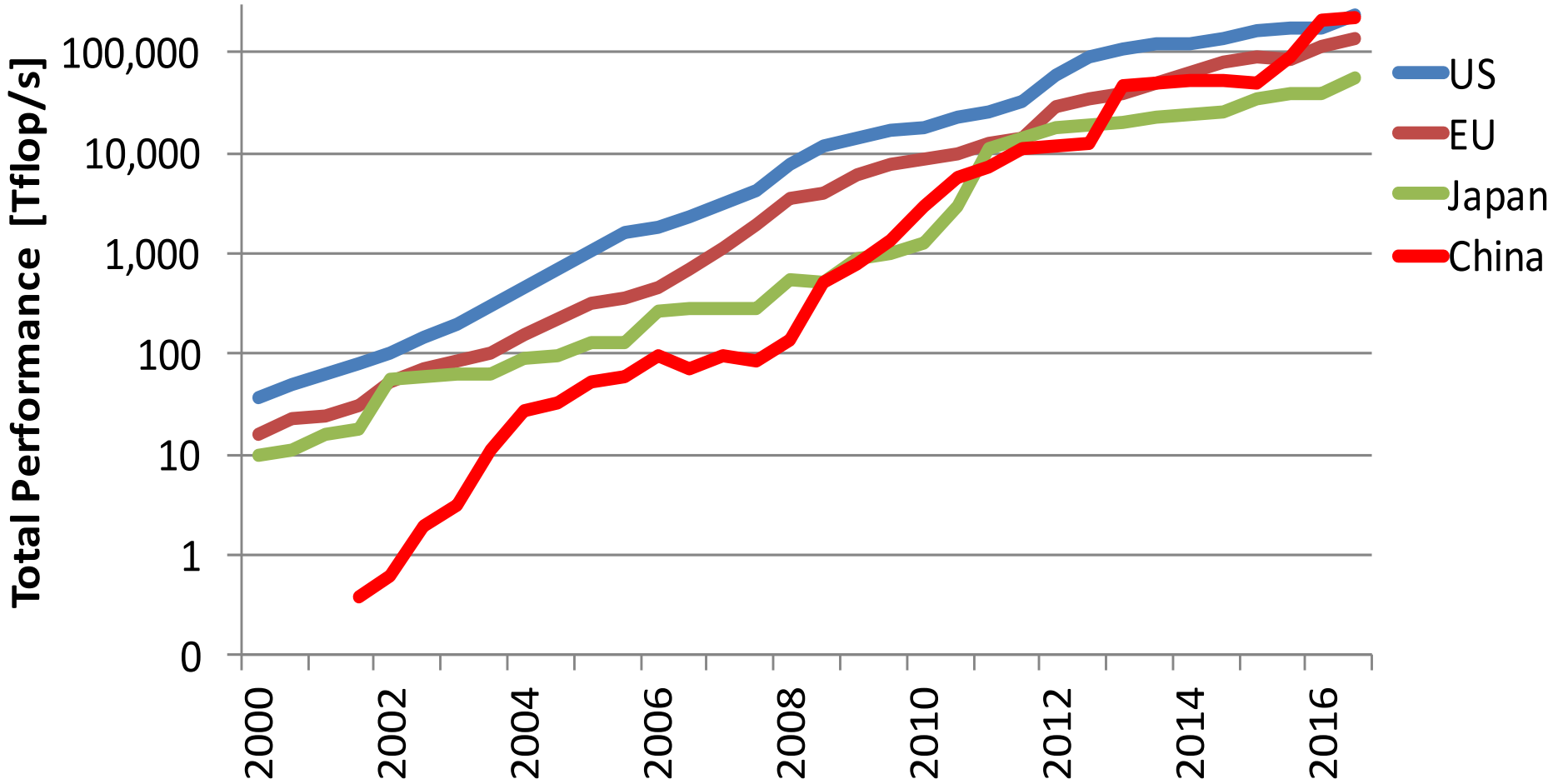
Input from: Erich Strohmaier, Patrick Naullieu (LBNL)
Satoshi Matsuoka (TiTech)
Haohuan Fu (Wuxi)
And many conversations in Singapore

John Shalf
Lawrence Berkeley National Laboratory

ASCAC, April 18, 2017

BERKELEY LAB
Lawrence Berkeley National Laboratory

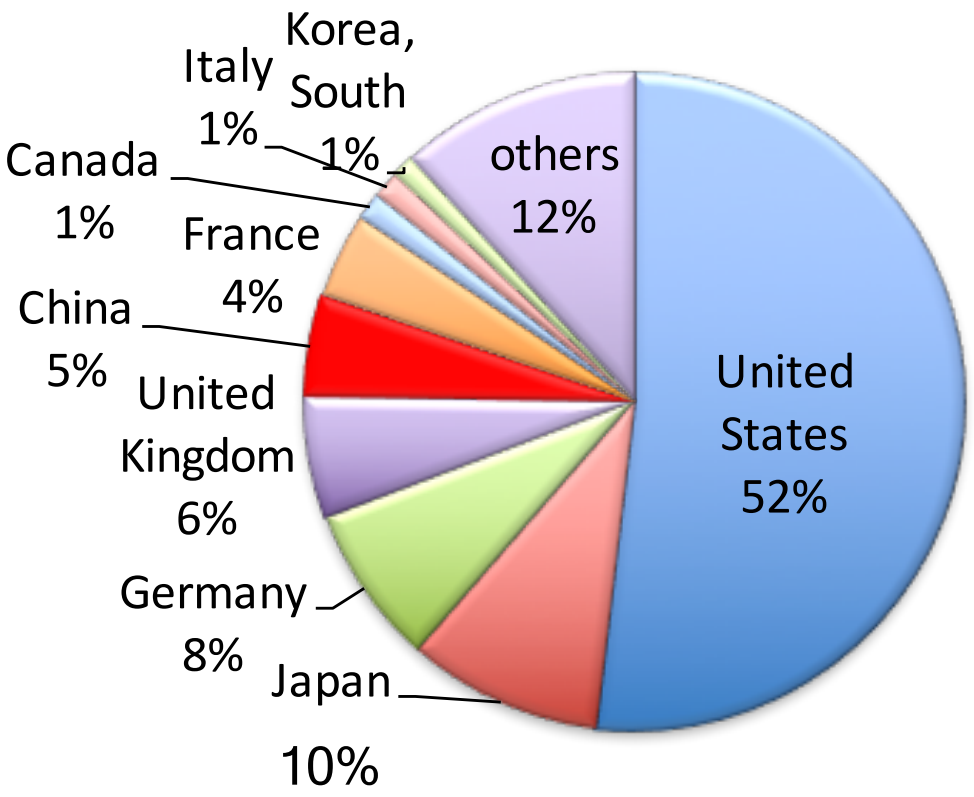
Performance of Countries



Share of Top500 Entries Per Country

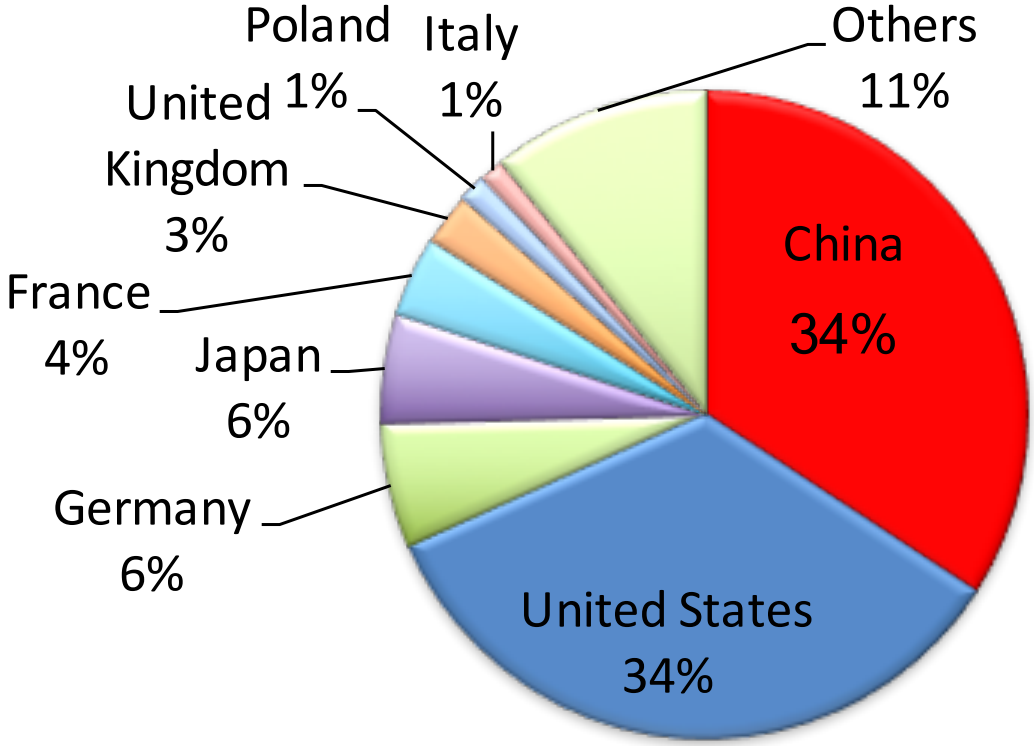
Historical Share

(averaged over liftetime of list)

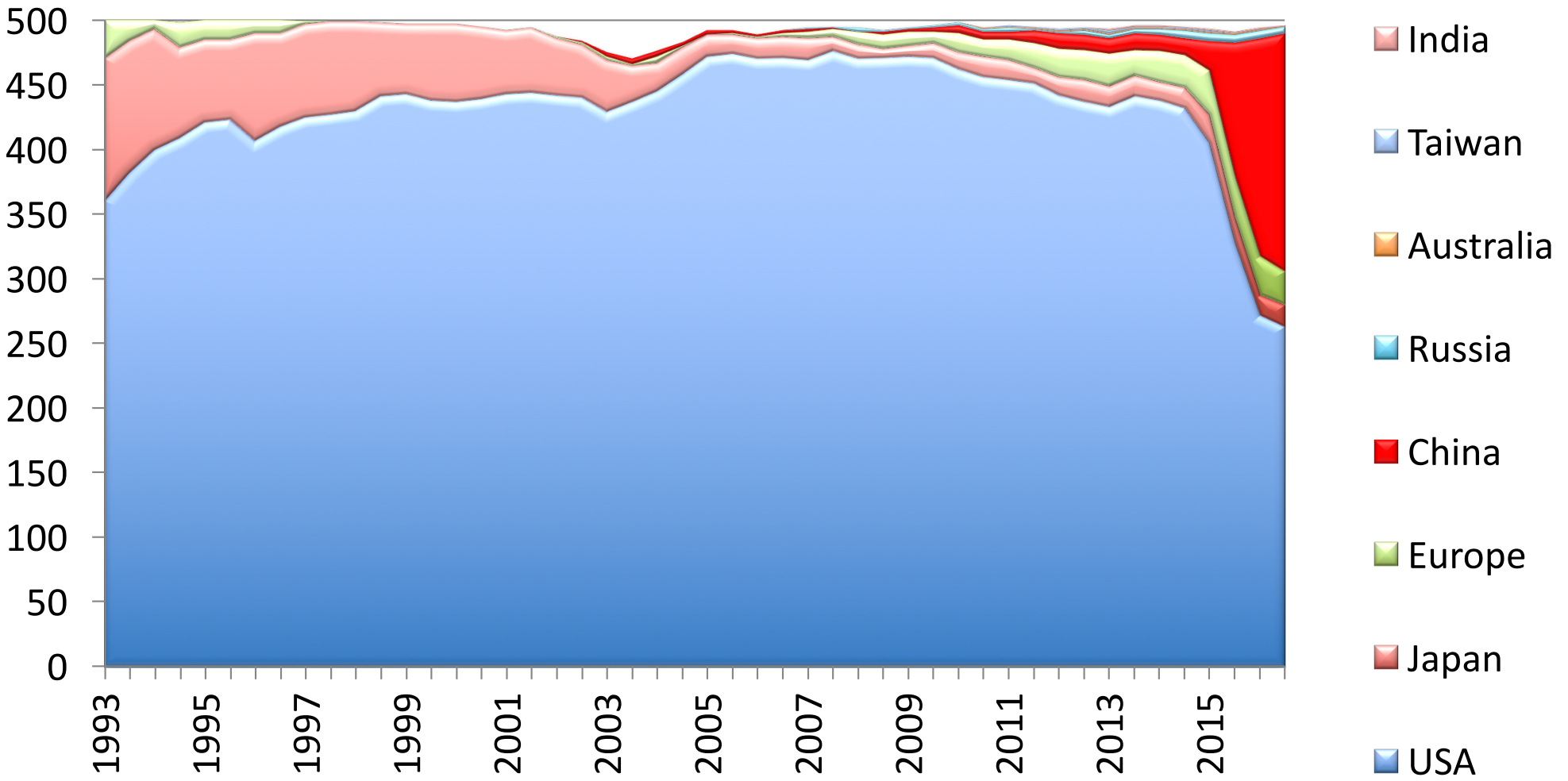


Current Share

(November 2016 list)

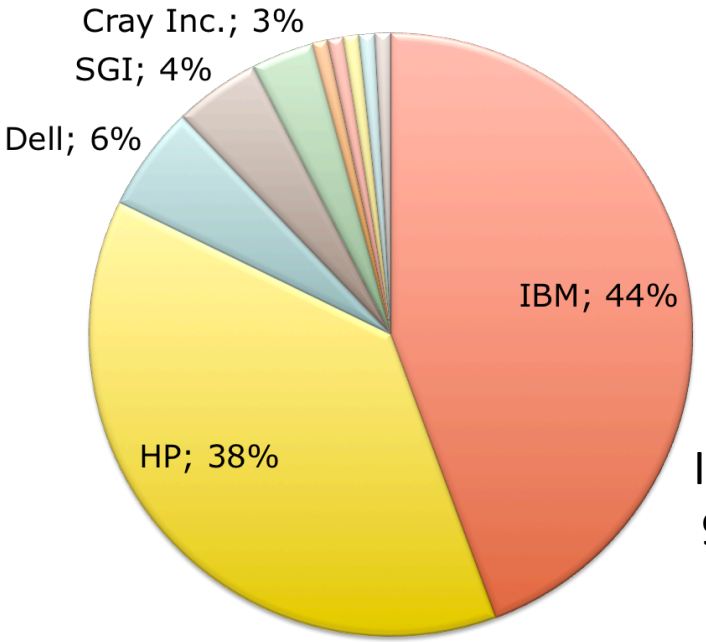


Producers of HPC Equipment

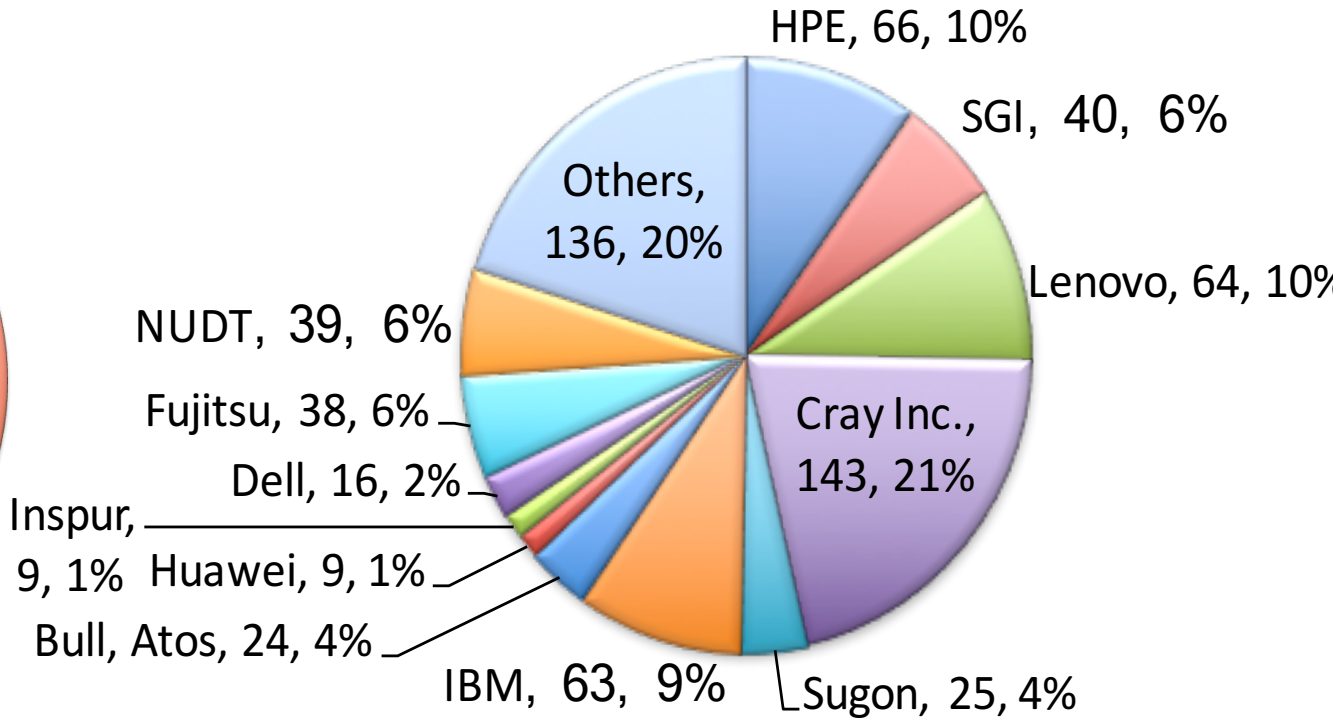


Vendors / Performance Share

2007



Now



Sum of Pflop/s, % of whole list by vendor

Top Level Conclusions

1. National security requires the best computing available, and loss of leadership in HPC will severely compromise our national security.
2. *HPC leadership has important economic benefits because of HPC's role as an enabling technology*
3. Leadership positions, once lost, are expensive to regain

Meeting participants expressed significant concern that – absent aggressive action by the U.S. – the U.S. will lose leadership and not control its own future in HPC

- It is critical to lead the exploration and development of innovative computing architectures that will unleash the creativity of the HPC community
- Workforce development is a major concern in HPC and a priority for supporting NSCI Objectives #4 and #5
- NSCI leadership develop more efficient contracting regulations to improve the public-private partnership in HPC science and technology development.

China Update

Aggressive Growth of China Chip Fabs

- **Current 28nm domestic capability in Shenzhen, Nanjing and other regions**

- **Broke ground on 14nm fab for 2018 near Shanghai**
  - Annual spending on fab equipment in China above $10B by 2018
  - Feb 2017: *China is expected to be the top spending region for fab equipment spending by 2019, overtaking South Korea and Taiwan.*

- **Foxconn offered 3T Yen ($30B) bid for Toshiba fabs**
  - Amazon & Google + SK Hynix & Western Digital consortium bidding
  - Apple bidding to own 20% stake in Fujitsu fab
  - TSMC withdrew its bid
  - Selection by June

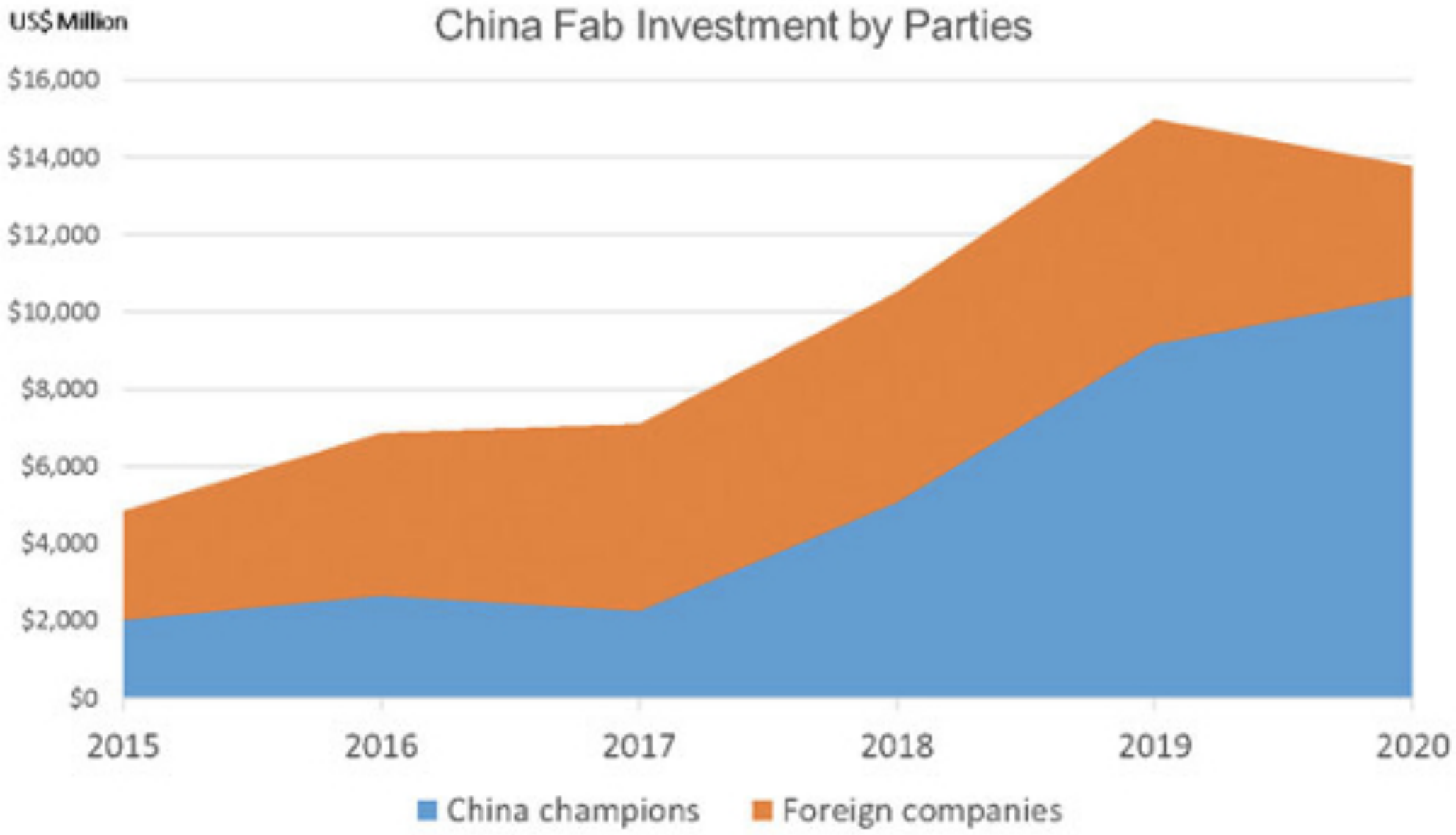
Fab Construction in China

Source Semiconductor Equipment and Materials International (SEMI)

Company	Location	Product	Begin Construction	Begin Production
Alpha & Omega	Chongqing	Power Discrete	tbd	tbd
Fujian Jin Hua	Fujian	DRAM	2016	2018
GigaDevice	Hefei	DRAM/Flash	tbd	tbd
GlobalFoundries	Chengdu	Foundry	2017	2018/2019
Hua Li Micro	Shanghai	Foundry	2016	2018
Powerchip	Hefei	Foundry	2015	2017
Samsung	Xian	3D NAND (Phase 2)	tbd	tbd
SMIC	Beijing Shanghai Shenzhen	Foundry Foundry Foundry	2016 2016 2016	2018 2018 2018
Tacoma Semiconductor	Nanjing	CMOS Image Sensor	tbd	tbd
Tsinghua Unigroup	Chengdu Nanjing	Foundry DRAM	tbd tbd	tbd tbd
TSMC	Nanjing	Foundry	2016	2018
UMC	Xiamen	Foundry	2015	2016
Yangtze River Memory/XMC	Wuhan	3D NAND	2016	tbd

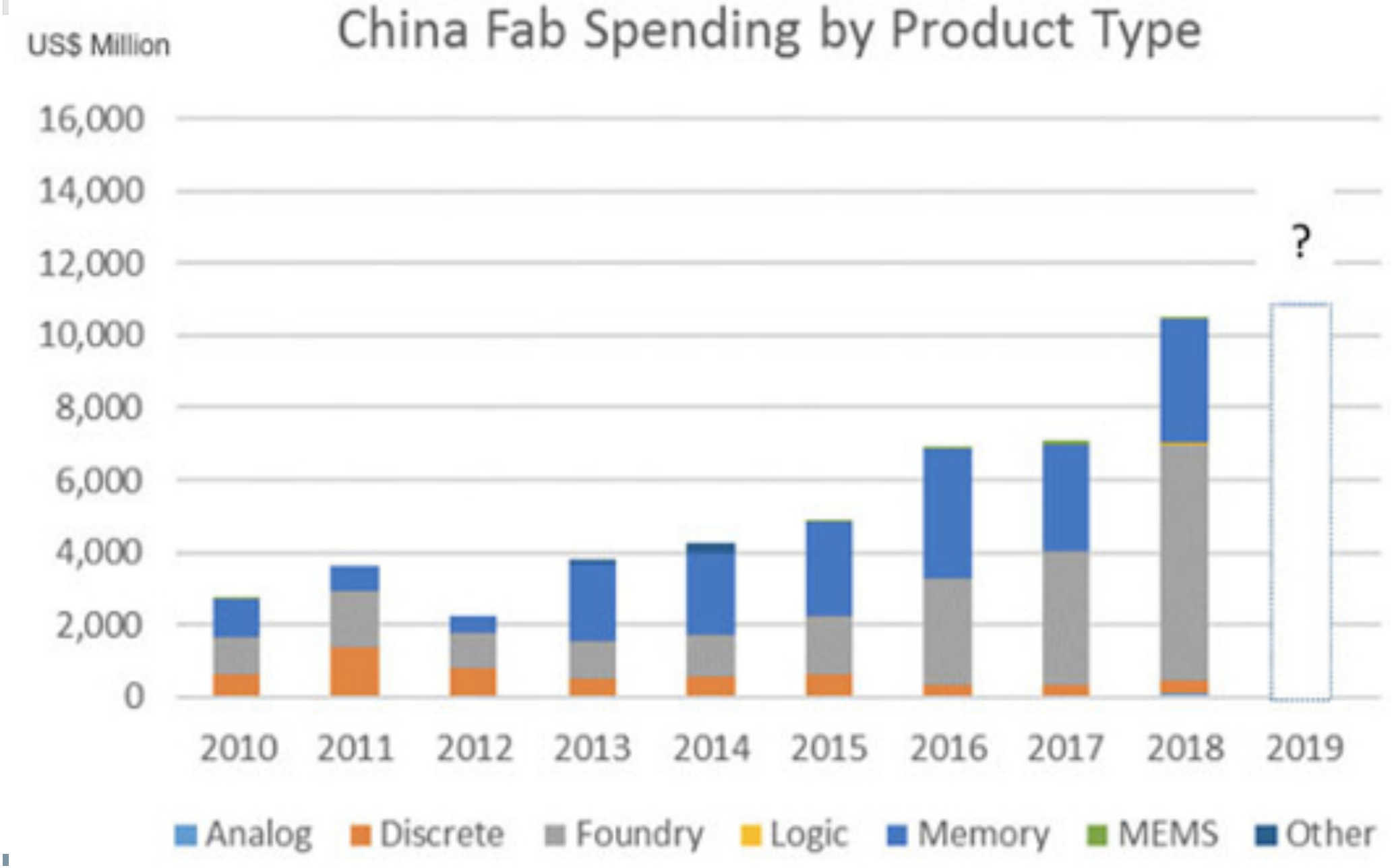
Fab Construction in China

Source Semiconductor Equipment and Materials International (SEMI)



Fab Construction in China

Source Semiconductor Equipment and Materials International (SEMI)

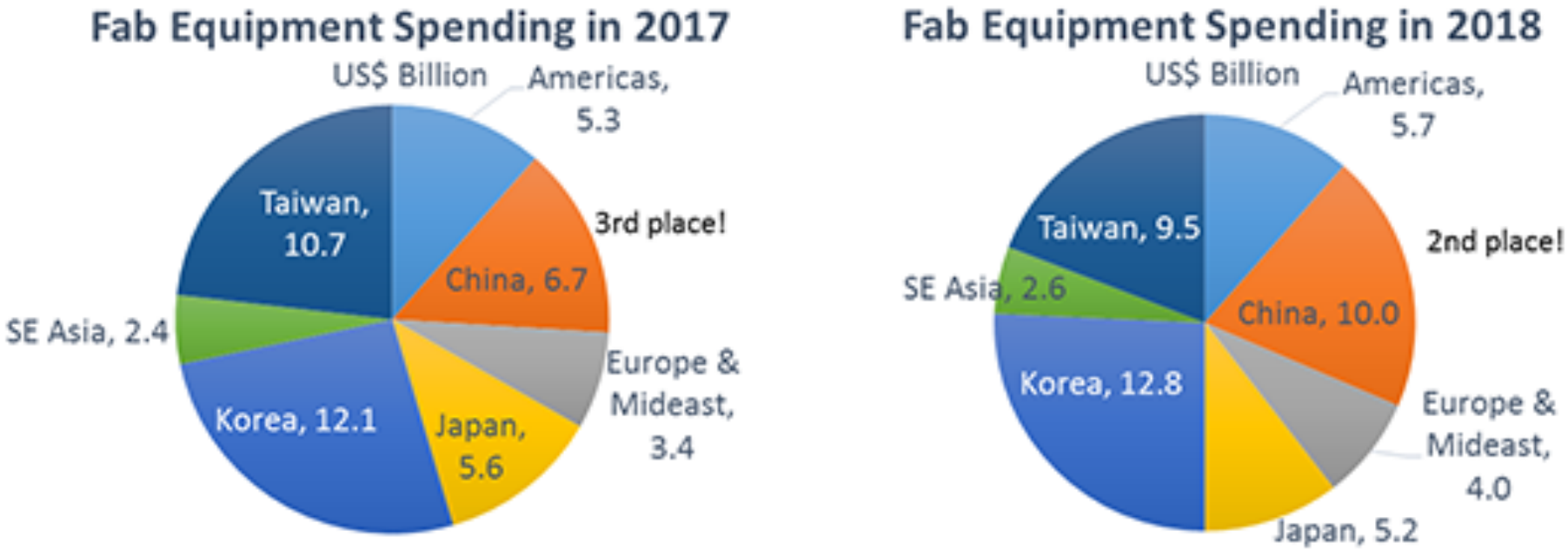


Fab Construction in China

Source Semiconductor Equipment and Materials International (SEMI)

Fab Equipment Spending by Region

Front End Facilities new & used (including Discretes & LED)



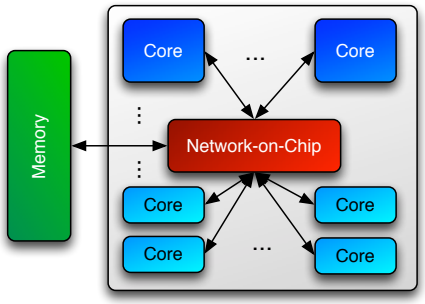
(World Fab Forecast report, Feb 2017, SEMI)

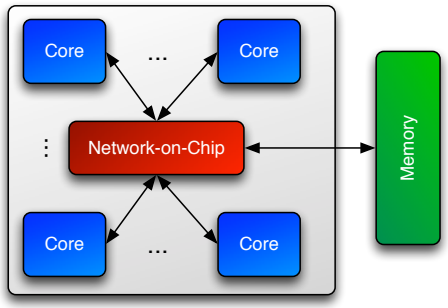
Source: World Fab Forecast report, SEMI August 2016

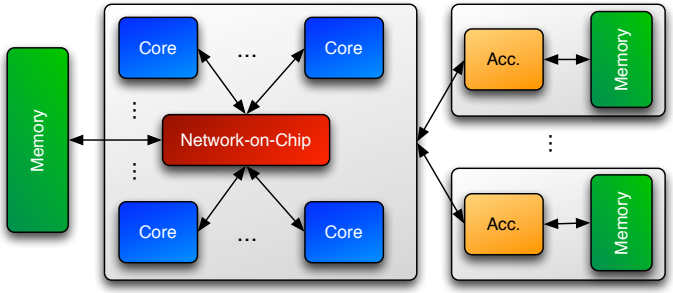
China 2017 Prototype System Bake-off

- **China plans to have three prototypes for candidate exascale systems delivered in 2017** [Xinhua: Jan 19, 2017]
- **Scale-up winner(s) to exascale in 2020** *(my guesses below)*
- *Other: Longsoon (unlikely), Silicon Cube (no), Thatic AMD (Tianjin/Sugon?)*

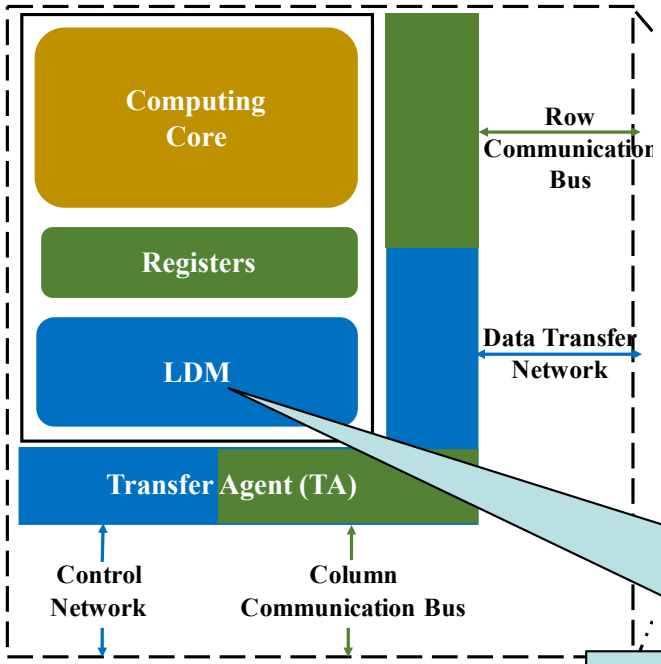
Wuxi/Sunway	NSC/Phytium	NUDT/Tianhe2a?
Heterogeneous manycore/accel	**Homogeneous Manycore**	**Attached Accelerator**
4*8x8 CPEs (light) + 4 MPE (heavy)	64-core ARMv8 self-hosted	ARMv8 PCIe attached accelerator (ISC16)
		<strategy may change>







Sunway Node Architecture
(refresher course)

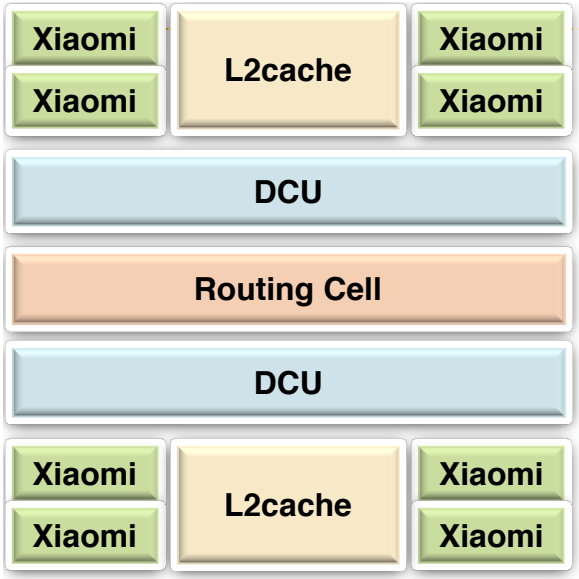


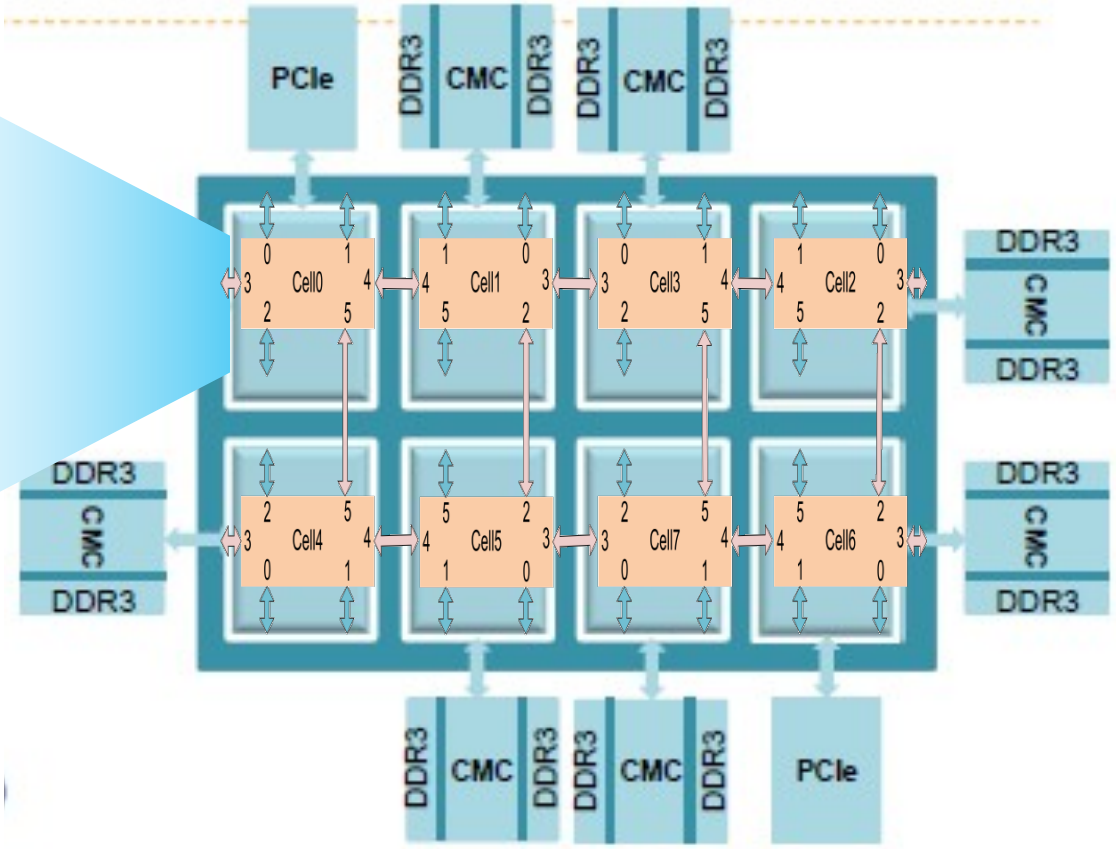
	DFMC (40 nm)
Architecture	4 CPE clusters (256 CPEs) 4 MPEs, 4 MCs
NoC	Mesh
On-chip memory	32 KB in each CPE×256=8 MB
Frequency	1 GHz
Computing ability	1000 GFLOPS DP
Memory bandwidth	102.4 GB/s DDR3
Chip area	$\sim$ 400 mm^2@40 nm
Power	$\sim$ 200 W

That is 64k per CPE LDM@28nm
(not 64k for the entire CPE mesh)
212 instructions Alpha-like ISA
240mm^2 chip area @ 28nm (cacti)

Fang Zheng (Wuxi)
Jan 2015
J. Comp. & Sci. Tech

Phytium Mars Architecture

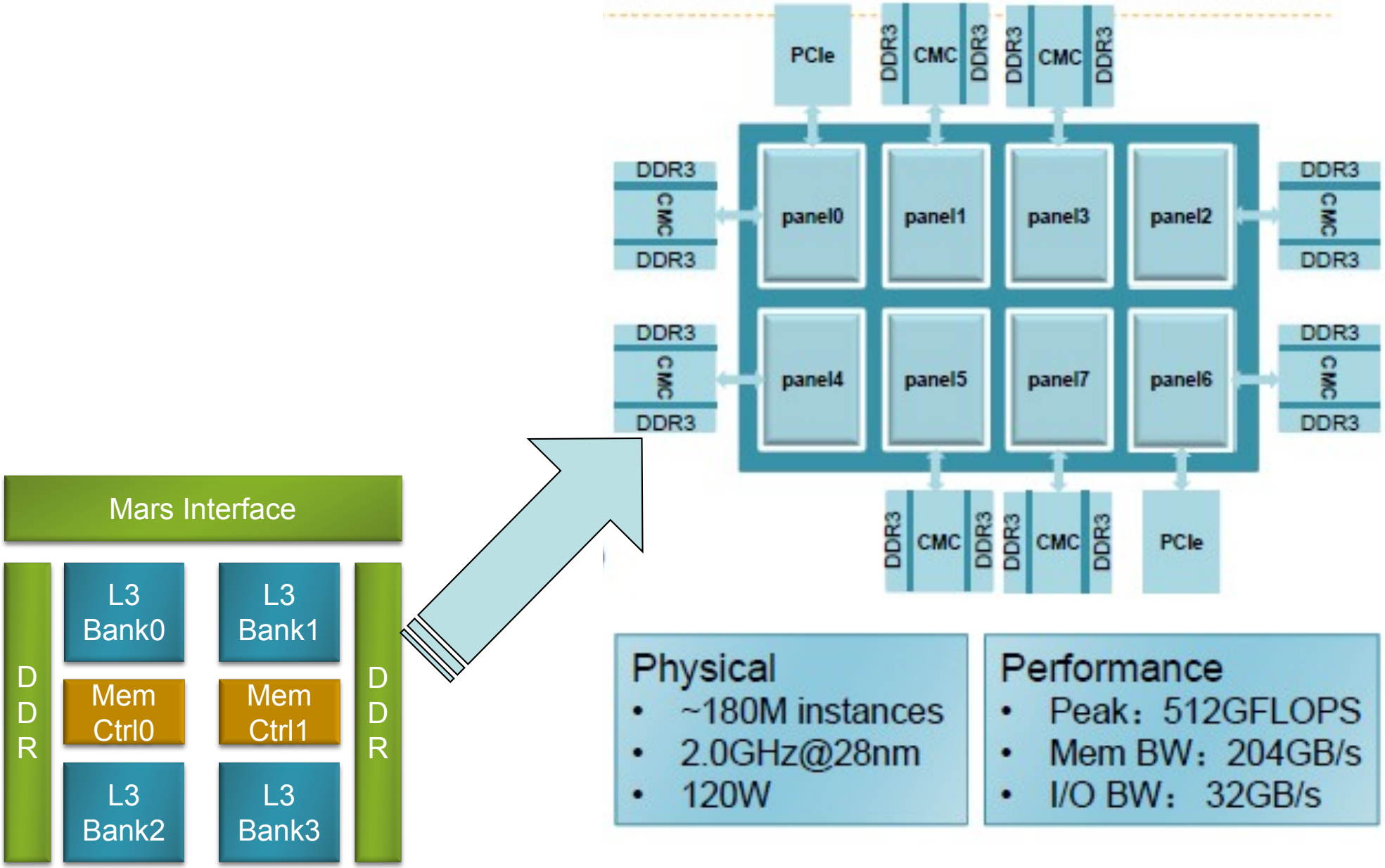




Physical
- ~180M instances
- 2.0GHz@28nm
- 120W

Performance
- Peak：512GFLOPS
- Mem BW：204GB/s
- I/O BW： 32GB/s

Phytium Mars Architecture

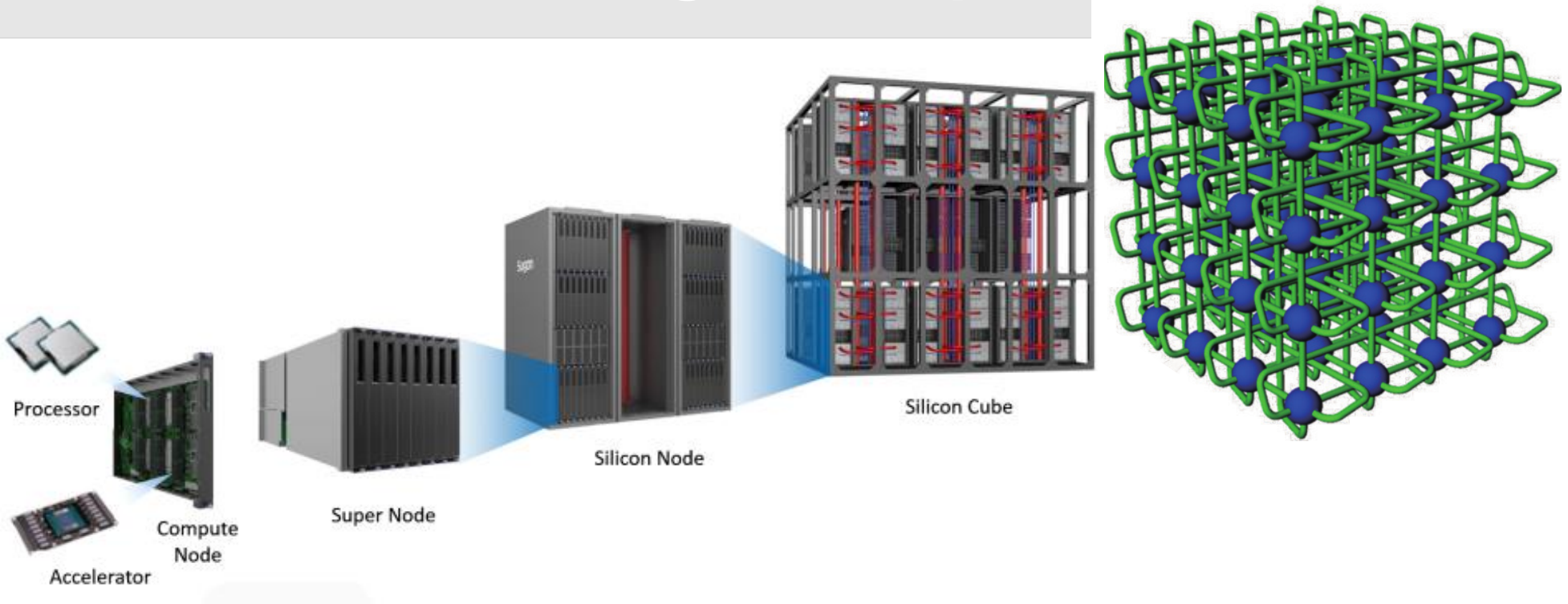


Comparing Phytium to Sunway

Category	Units	Phytium/Mars	Sunway	Ratio
ISA	-	ARMv8	CPE (DSP-like)	
Cores/numanode	cores	64	64	1.0
Core FLOP Rate	GFLOPs	8	11.72	0.7
L1$/Core	KB	32	64	0.5
Clock Rate	GHz	2GHz	-	-
Power/numanode	W	120	93	1.3
Performance/numanode	GFLOPs	512	750	0.7
Memory Bandwidth/numanode	GB/s	204	34	6.0
Sockets for 125 PF system	-	234,375	40,960	5.7
Cores for 125PF system	Millions	15	20	0.8
Power for 125 PF system	MW	28	15	1.9

- **Phytium advantages**
  - 6x higher memory bandwidth per NUMA node
  - Conventional CPU programming model
- **Sunway advantages**
  - 2x Energy efficiency of Phytium system
  - 5x higher performance density (5x fewer sockets for a system)

Sugon Silicon Cube: Meteorological Supercomputer



- **Processor**
  - Intel Xeon E5-2680 12core
  - DDR4 2133MHz memory
- **Interconnect**
  - FDR InfiniBand (56Gb)
  - Torus interconnect topology

- **Overall System**
  - Peak 1 PFLOPS
  - #95 non Top500 at 75% efficiency
  - Total Memory Capacity 80TB
  - 208 square meters of 1,000 liquid cooled servers
  - Total power is 641.38kW

What's in a Name? Sunway

Shen "God"

Wei "Powerful"

神威

太湖之光

Taihu Lake
A famous lake near Shanghai

apostrophe
"Taihu's"

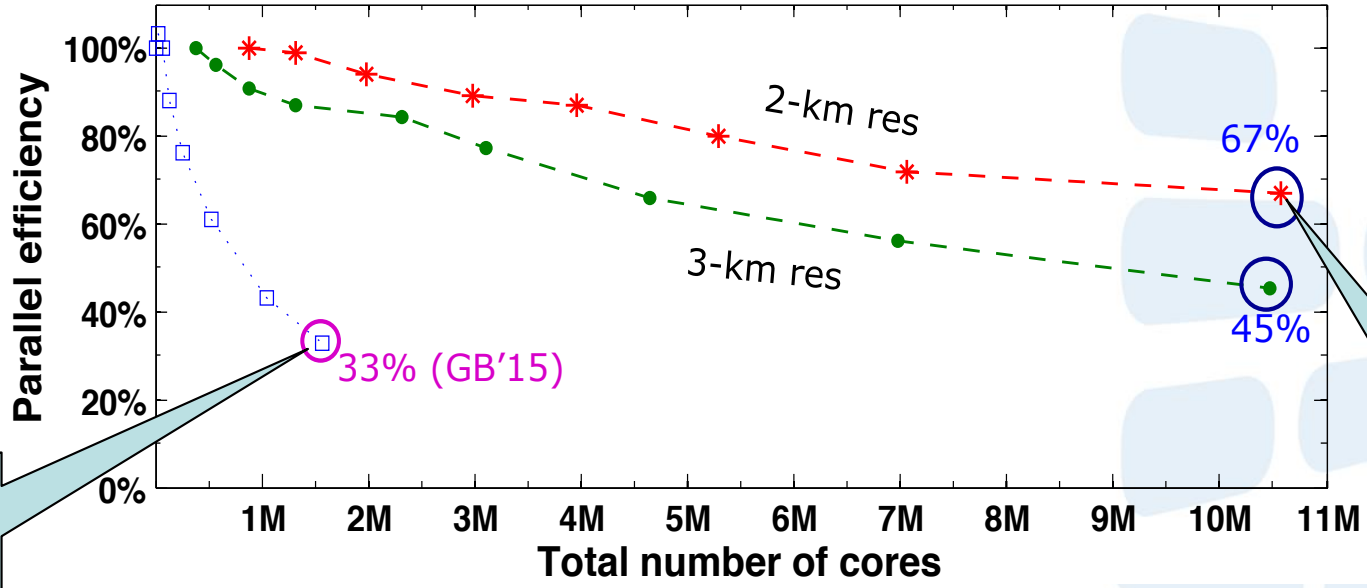
Guang
"Light"

System Comparisons

System	TaihuLight	Tianhe-2	Titan	Sequoia	Cori
Peak Performance (PFlops)	**125.4**	54.9	27.1	20.1	27.9
Total Memory (TB)	**1310**	1024	710	1572	879
Linpack Performance (PFlops)	**93.0(74%)**	33.9(62%)	17.6(65%)	17.2(85.3)	14.0(50%)
Rank of Top500	**1**	2	3	4	5
Performance/Power (Mflops/W)	**6051.3**	1901.5	2142.8	2176.6	3266.8
Rank of Green500	**4**	135	100	90	26
GTEPS	**23755.7**	2061.48	###	23751	###
Rank of Graph500	**2**	8	###	3	###
HPCG (Pflops)	**0.3712**	0.5801	0.3223	0.3304	0.3554
Rank of HPCG	**4**	2	7	6	5

Continued Progress Since GB Runs

(Wuxi team are committed to codesign)



Haohuan Fu
Wuxi

Initial Port of the CCSM3 model

2016-2017 improvement through sync-free data-locality preserving algorithm

cube-sphere grid or other grid

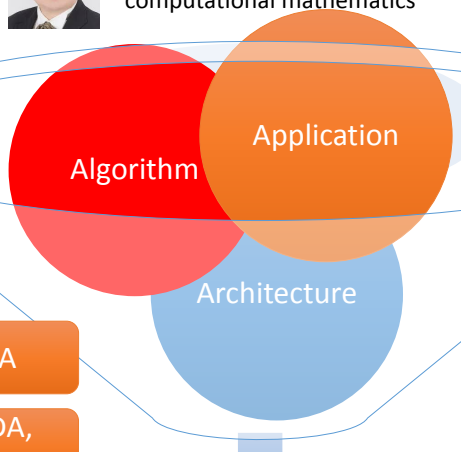
explicit, implicit, or semi-implicit method

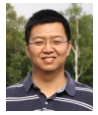
Yang, Chao
Institute of Software, CAS
computational mathematics

cloud resolving

Wang, Lanning
Beijing Normal Universit
climate modeling

Xue, Wei
Tsinghua University
computer science





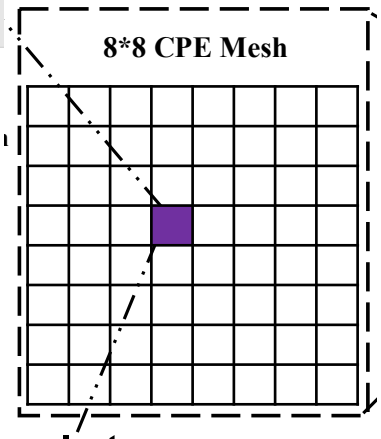
Fu, Haohuan
Tsinghua University
geo-computing

Sunway, GPU, MIC, FPGA

C/C++, Fortran, MPI, CUDA, Java, …

The "Best" Computational Solution

Programming Model MPI+X



- **MPI3**
  - Based on OSU MVAPICH
  - One MPI process per MPE core (4 per socket)
- **OpenACC 2.0**
  - OpenACC2.0 cross-compiler based on LLNL ROSE translator
  - Fortran, C/C++ Support
  - Nearly identical accelerator offload style as for GPU systems
  - Copy in to “fast memory” is CPE local stores instead of GPU GDDR5 memory.
  - Extensions (*swap,pack,tilemask*) for hardware collective mem ops
- **Athreads:** A low-level spatial threading
  - Low-level target of the ROSE OpenACC translator
  - Supports some hardware collective operations such as transposes and common domain decomposition operations (beyond Pthreads)

Comparison of OpenACC Offload Models

From Haohuan Fu
Wuxi

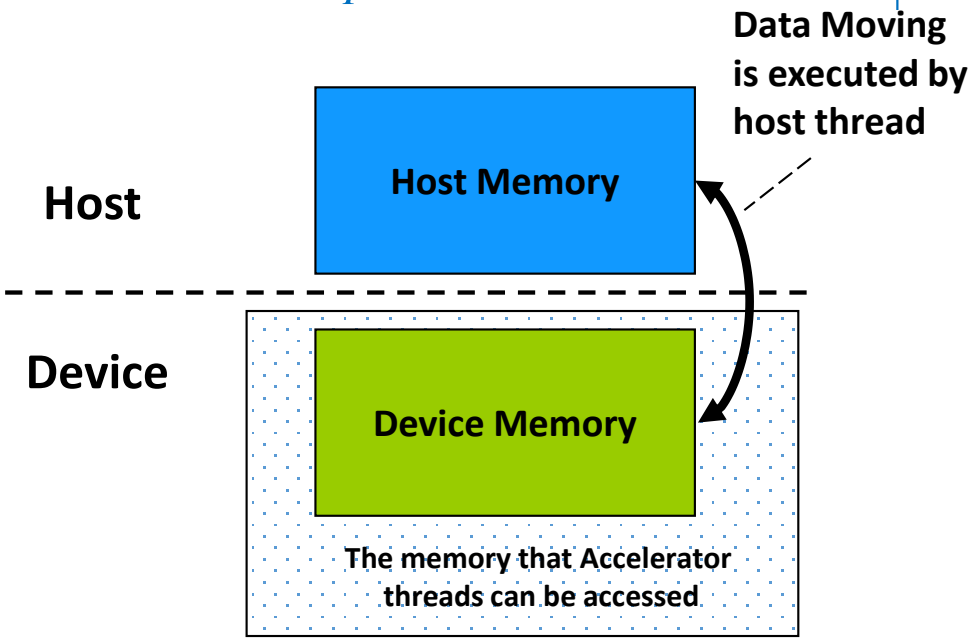
- Use *data copy* to handle data moving between Mem and LDM

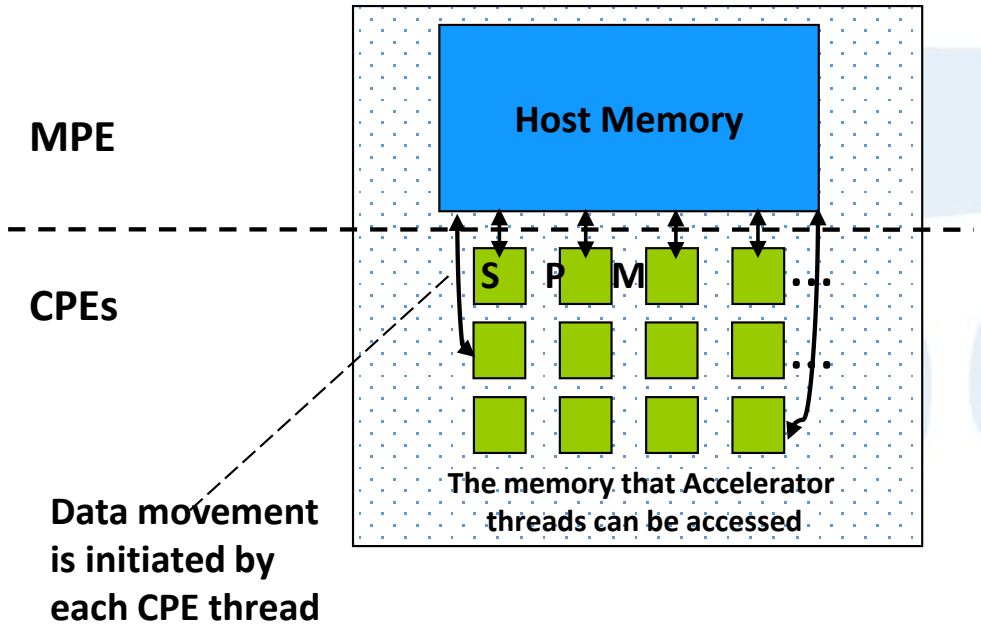
```
!$acc data copyin(A) copyout(B)
!$acc parallel loop
  do i=1,128
     m = func(i)
      do j=1,128
        B(j, i) = A(j, m)
     enddo
  enddo
  !$acc end parallel loop
  !$acc end data
```

OpenACC2.0

```
!$acc parallel loop
do i=1,128
   m = func(i)
    !$acc data copyin(A(*, m)) copyout(B(*, i))
   do j=1,128
      B(j, i) = A(j, m)
   enddo
    !$acc end data
enddo
!$acc end parallel loop
```

Sunway OpenACC





Software Porting Strategy

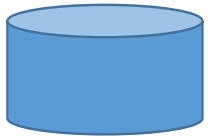
The Gap between Software and Hardware

From Haohuan Fu
Wuxi

100P

- millions lines of legacy code
- poor scalability
- written for multi-core, rather than many-core

100T



China's models
- pure CPU code
- scaling to hundreds or thousands of cores

China's supercomputers
- heterogeneous systems with many-core chips
- millions of cores

Software Porting Strategy

Our Research Goals

From Haohuan Fu
Wuxi

- highly scalable framework that can efficiently utilize many-core processors
- automated tools to deal with the legacy code

100P

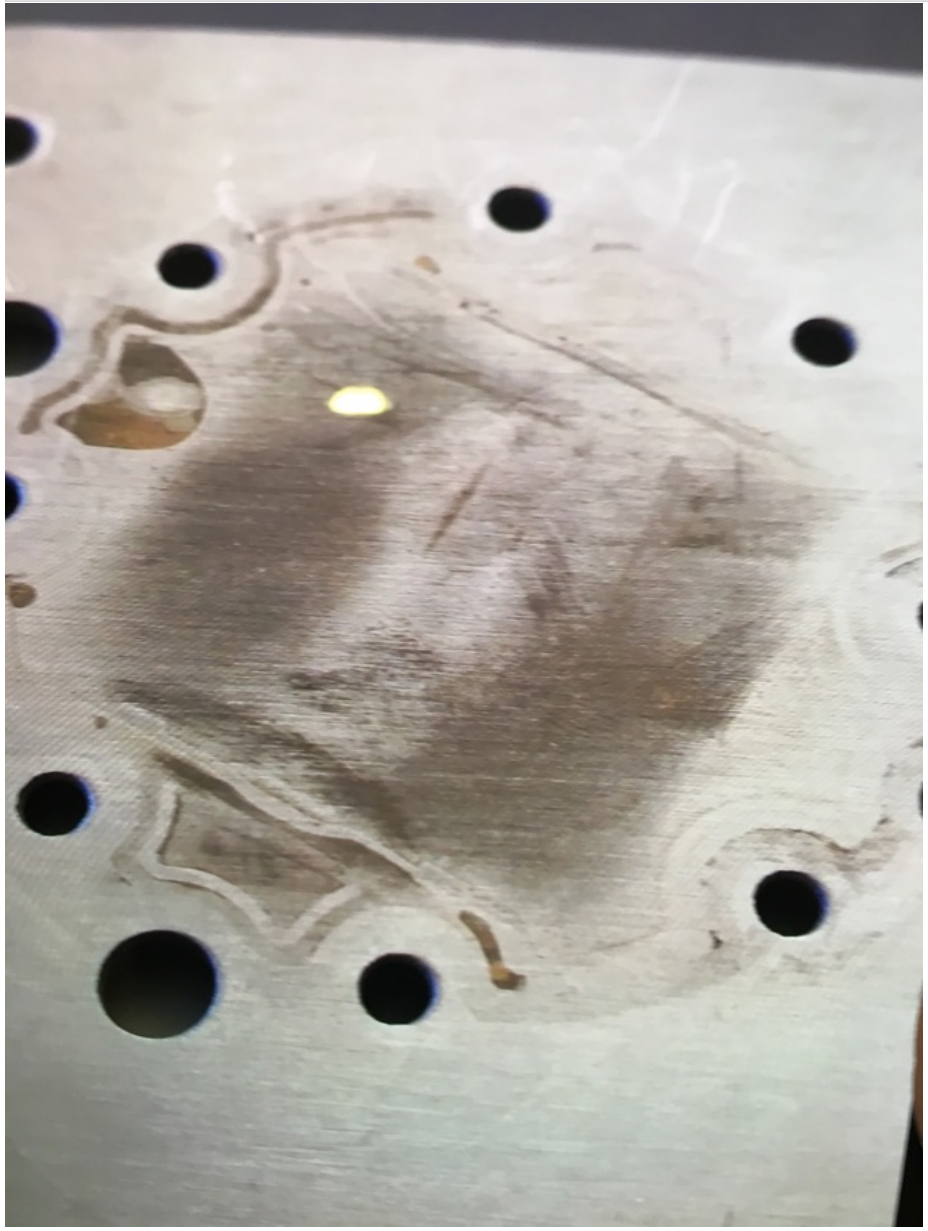
- millions lines of legacy code
- poor scalability
- written for m[illegible]e, rather than many-core

100T

China's models
- pure CPU code
- scaling to hundreds or thousands of cores

China's supercomputers
- heterogeneous systems with many-core chips
- millions of cores

Is this and image of “failure” or of “success”?



Can anyone guess what this is?

- **Apple Ethos**: Refine until it is near perfect.
- **Google Ethos:** Try early and try often!

Conclusions on China

Hardware Strategy

- 3 Prototype Systems in 2017 (Sunway, Phytium, ?Tianhe2a?)
- ARMv8 systems more conventional than Sunway (less energy efficient)

Software Strategy

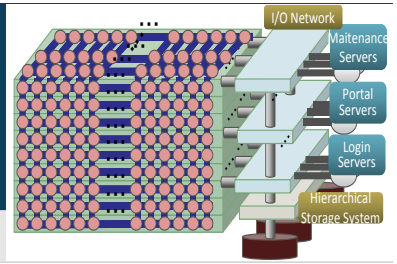
- MPI+x where x=OpenACC directives
- Sunway OpenACC programming similar to GPU systems (not exotic)
- Plans to increase automation to port from old to new
- Continue advances in algorithm design increase gains over GB wins

Overall: Moving at a fast pace

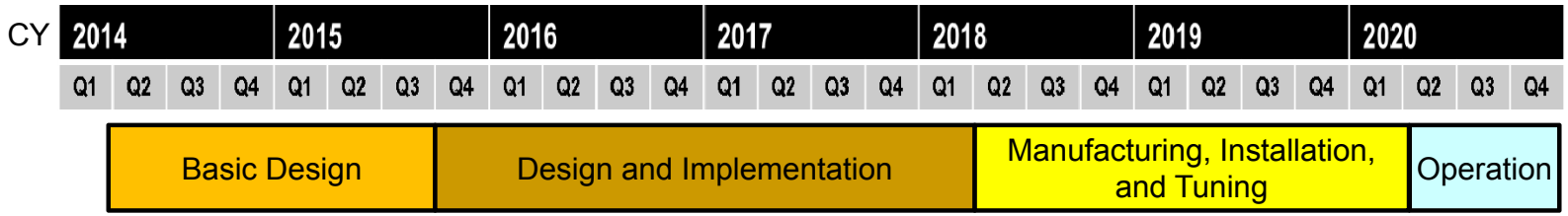
- Investing in a portfolio of risk (*ranging from conventional to exotic*)
- There is little incentive to play it safe (no *alternatives*)
- And are not held back by an installed base *(open ended design space)*
- **Not about out-selling US in HPC!** *Its about creating domestic supply chain to support domestic industry (cars, aerospace, basic science…)*

Japan Update

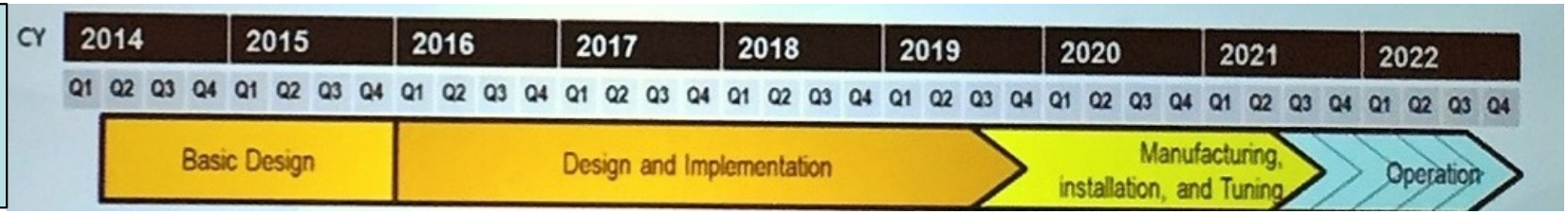
Post-K Strategic Delay



Old Timeline



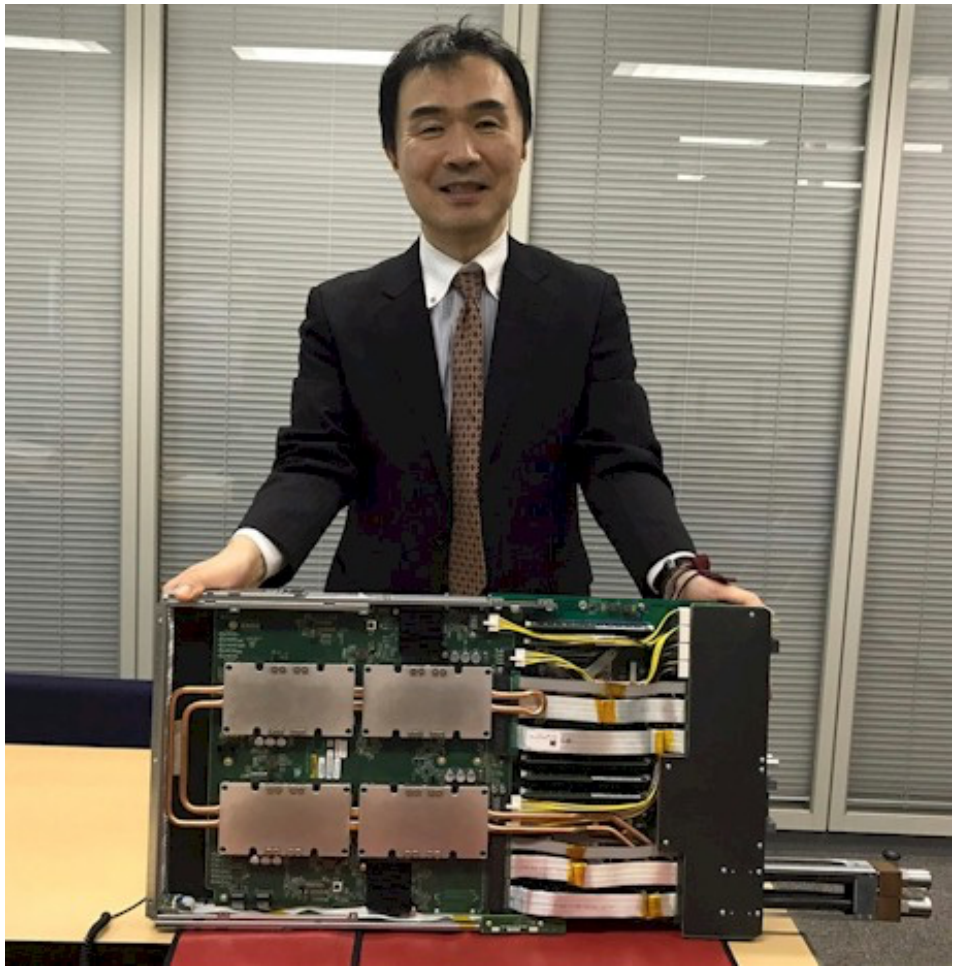
New Timeline

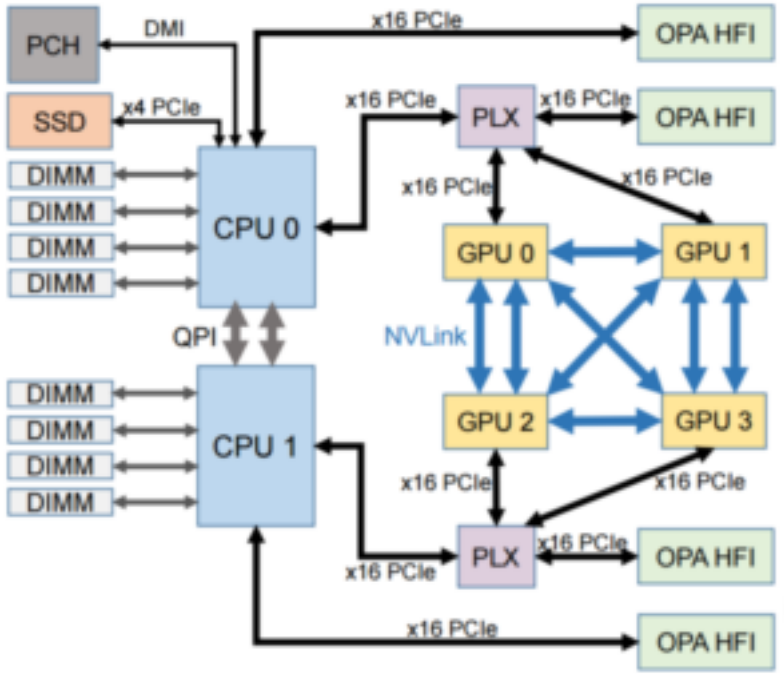


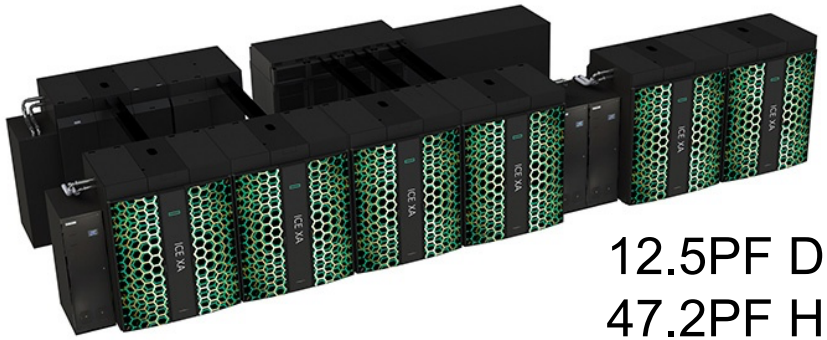
- **Flasgship 2020: 0.91B USD project (Riken + Fujitsu)**
- **Originally planned for 2020 → moving to 2021 or 2022**
  - Energy efficiency benefits of new process technology offer better TCO
- **Fujitsu: scalable-core/node ARMv8 + SVE512 Vectors**
  - **Wider vectors:** K=128, FX100 256, Post-K 512
  - *6D Mesh Interconnect, Sector Cache, Fast Sync between cores*
  - ***Nearly same microarchitecture as SPARC64-based K-computer***
  - *Gains advantage of larger market for ARM software ecosystem*

Tsubame 3.0 @ TiTech/SGI/HP/NVIDIA

Converged BigData/AI/HPC Supercomputer







12.5PF DP
47.2PF HP
Omnipath@ 4x100Gb/s

ML moving towards AI is a Hotbed of Activity in US & Japan HPC

Figure: Fujitsu

ARTIFICIAL INTELLIGENCE

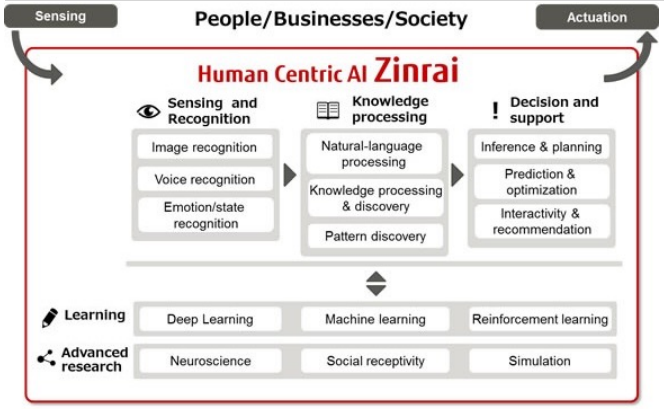
A program that can sense, reasons, act and adapt

MACHINE LEARNING

Algorithms whose performance improve when exposed to more data over time

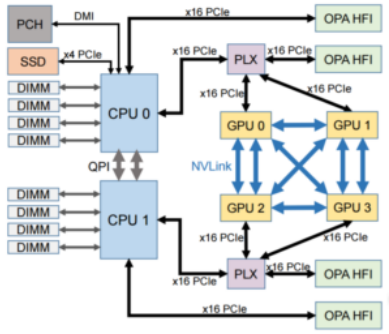
DEEP LEARNING

Multi-layered neural networks learn from vast amounts of data

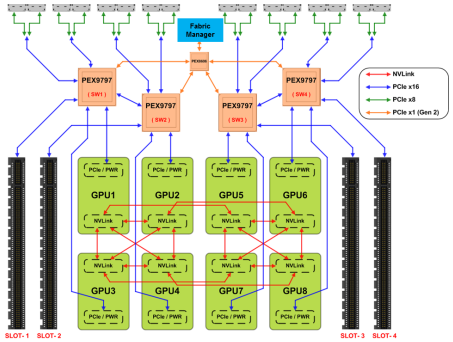


?

Tsubame 3.0



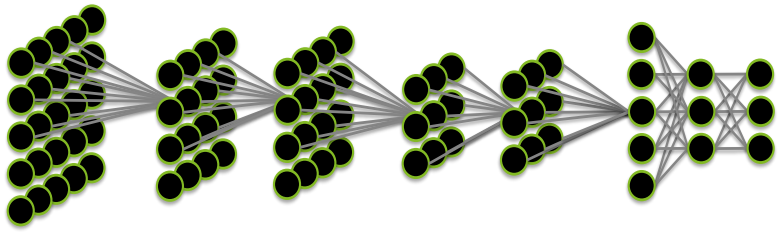
MV DGX-1 & Fujitsu

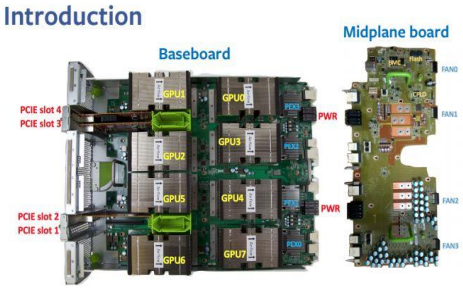


Google TPU
MS Olympus

FACEBOOK BIG BASIN

8x Tesla P100 GPU Server - Hybrid Mesh Cube Topolog

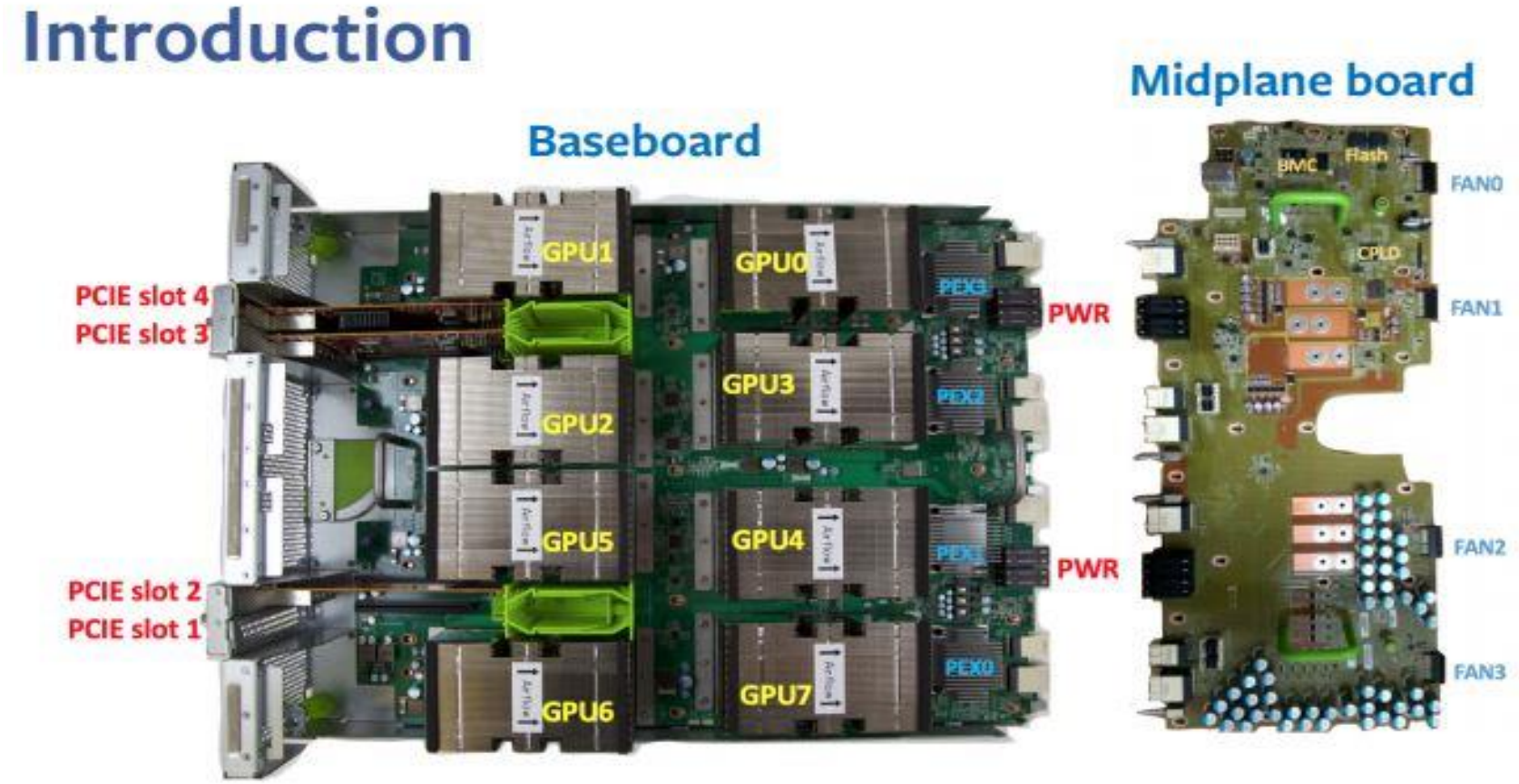




Vertically Integrated Data Centers Spinning their Own Designs

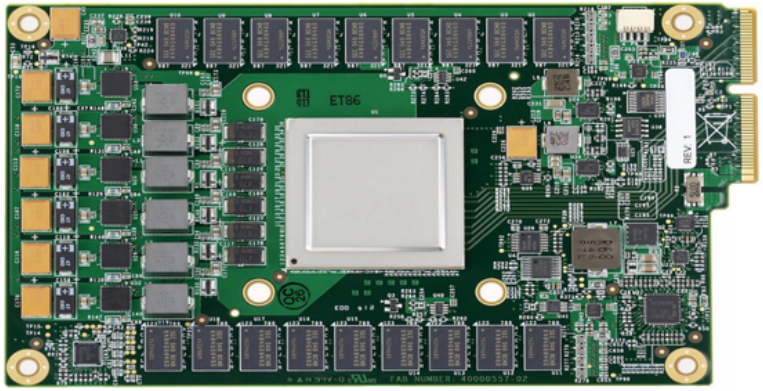
FACEBOOK BIG BASIN

8x Tesla P100 GPU Server – Hybrid Mesh Cube Topology



Google Tensor Processing Unit (TPU)

- **Deployed in datacenters since 2015**
- 10-30x Faster than NVIDIA G80 or Intel Haswell for ML workloads *(64k arithmetic ops per cycle)*
- Could be faster with better memory subsystem.
- 8bit integer arithmetic (all that is needed for ML)



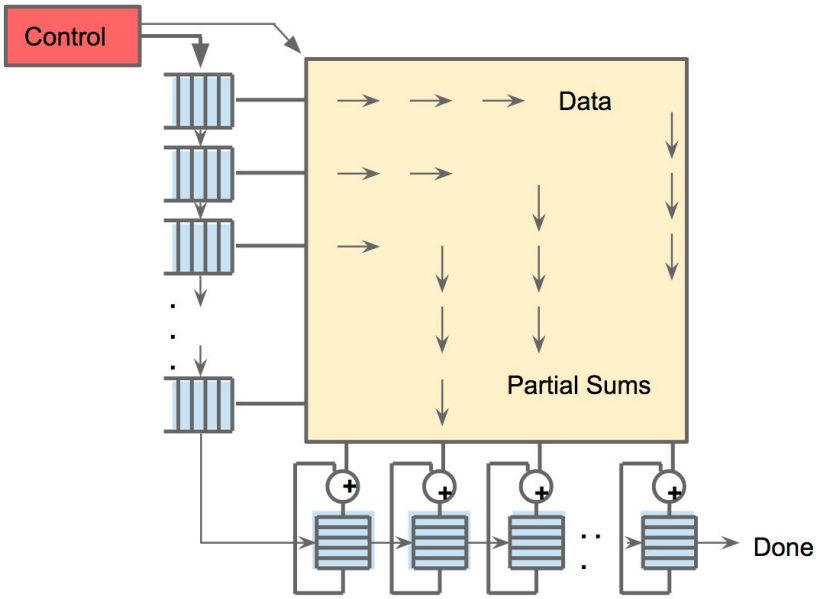
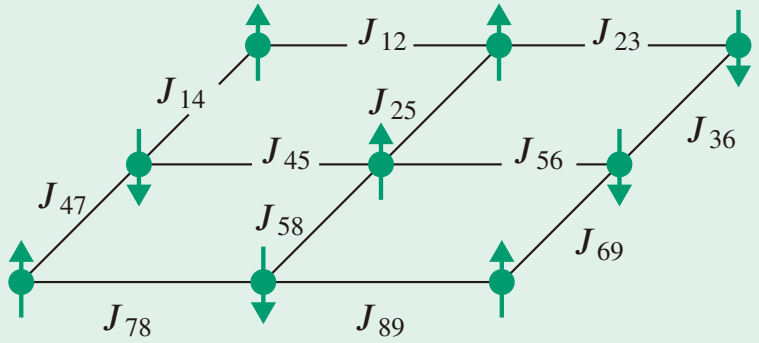


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

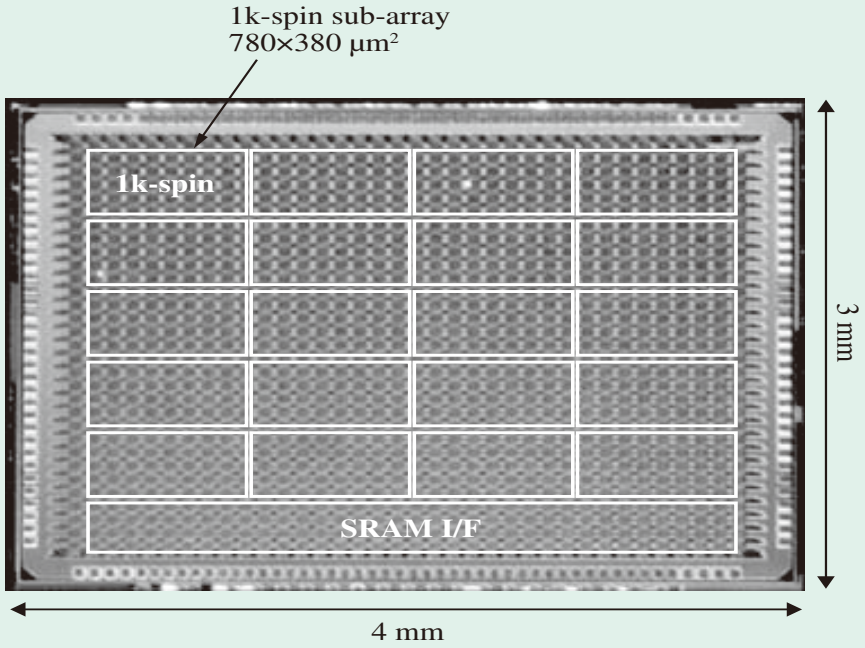
Model	Die										Benchmarked Servers				
	mm^2	nm	MHz	TDP	Measured		TOPS/s		GB/s	On-Chip Memory	Dies	DRAM Size	TDP	Measured	
					Idle	Busy	8b	FP						Idle	Busy
Haswell E5-2699 v3	662	22	2300	145W	41W	145W	2.6	1.3	51	51 MiB	2	256 GiB	504W	159W	455W
NVIDIA K80 (2 dies/card)	561	28	560	150W	25W	98W	--	2.8	160	8 MiB	8	256 GiB (host) + 12 GiB x 8	1838W	357W	991W
TPU	NA*	28	700	75W	28W	40W	92	--	34	28 MiB	4	256 GiB (host) + 8 GiB x 4	861W	290W	384W

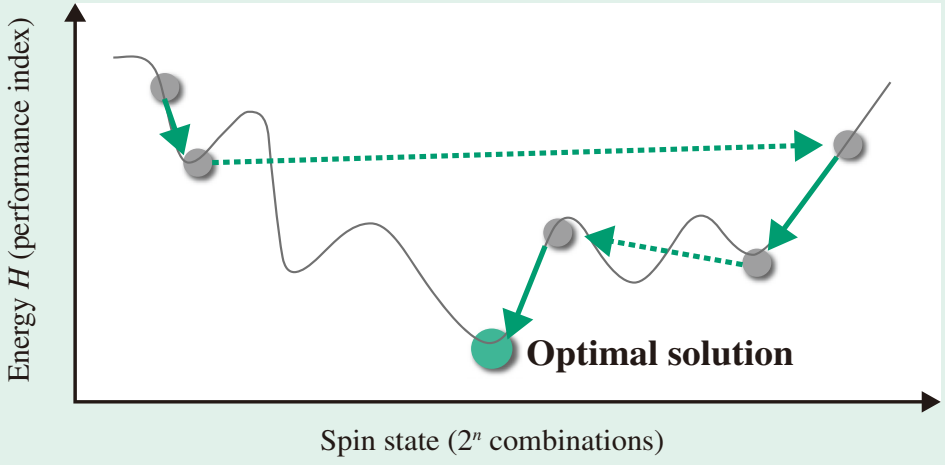
Hitachi 20k spin "Ising Chip"

Similar to D-Wave quantum annealer, but room temp.



$$H = -\sum_{\langle i,j \rangle} J_{ij}\,\sigma_i\,\sigma_j - \sum_i h_i\,\sigma_i$$





	New technique	Existing technique
Approach	Ising computing	
	Semiconductor (CMOS)	Superconductor
Operating temperature	Room temperature	20 mK
Power consumption	0.05 W	15,000 W (including cooling)
Scalability (number of spins)	20,000 (65 nm) Can be scaled up by using higher level of scaling	512
Computation time	Milliseconds	Milliseconds (fast in principle)

A Brief EU Update

Recent Developments in the EU
(and soon to be former members thereof)

- **Jan 18 2017: "***Cray commits to deliver 10,000+core ARM system"*
  - **"Isambard"** @ UK Bristol £4.7M to be installed March-December 2017
  - Includes GPUs, x86 CPUs, and FPGAs (in addition to ARM)
  - **Simon McIntosh-Smith:** *"Scientists have a growing choice of potential computer architectures to choose from, including new 64-bit ARM CPUs, graphics processors, and many-core CPUs from Intel. Choosing the best architecture for an application can be a difficult task, so the new Isambard GW4 Tier 2 HPC service aims to provide access to a wide range of the most promising emerging architectures, all using the same software stack."*
- **Change in EU Horizon 2020 strategy in Feb 2017**
  1. Refocus on domestic technologies
  2. Preparatory call for proposals expected imminently
  3. Open to non-traditional architectures (e.g. EU BRAIN project)
  4. Current Focus on ARM (ISA license, but indigeonous microarchitecture)
  5. Chiplet and SoC integration strategies are both being pursued

Conclusions

- **China to have 3 prototype systems by 2017** (*exascale candidates*)
  - Try early, try often, learning cycle
  - Broad range of architectures.
  - Not constrained by an installed base. (*both an asset and a curse!*)
- **China's Sunway system has emboldened other countries to pursue an "all indigenous" processor approach**
  - Enabled by embedded ecosystem (*don't have to own "all" of the design*)
  - Started with China and Japan, but now EU has joined in to the strategy
- **Japan refocusing Flagship 2020**
  - New roadmap for ARMv8 based Post-K (2021-2022)
  - Innovations happening at smaller scale for ML acceleration (TiTech)
- **ML is taking off in Asia and US**
  - Plus: Driving a lot of innovation and investment in HPC-relevant technologies (*contributions vertically integrated Inc. Google/FaceBook*)
  - Minus: Focus is on low precision arithmetic (8 bits floats?!?!)
  - Broader trend towards AI (a superset of ML and neural networks)

Unofficial Intel timeline: 7nm chips in 2020, 5nm in 2022?

01/21/2016 at 10:32 AM by Brad Linder

> IP Transit $0.20/Mbps
> 17,000 BGP sessions with over 5,700 Networks plus 149 Internet Exchange Points.

Intel used to release chips on a Tick-Tock schedule, which basically meant that use a new manufacturing process. Since 2006 Intel moved from 65nm chips to 14nm chips in this fashion.

Intel is pouring $7 billion into 7nm chip production plant in Arizona

By Paul Lilly 2 days ago

Looking beyond Cannonlake.

Samsung Galaxy S9 to use new 7nm chips in 2018 – report

0 shares Read Comments

Report: TSMC lays the groundwork for a 5-nm and 3-nm foundry

by Wayne Manion — 9:28 AM on December 12, 2016

The chip-making world may be gearing up for 10-nm chip manufacturing processes, but according to a report by the Nikkei Asian Review, contract silicon manufacturing house Taiwan Semiconductor Manufacturing Company (TSMC) is getting ready to build a fab for what it calls "5-nm" and "3-nm" process nodes. The company reportedly expects the plant to cost approximately $500 billion New Taiwan dollars, an amount equivalent to $15.8 billion U.S. dollars. According to the reports, TSMC has already asked the Taiwanese government for assistance in finding a location of sufficient size for the new factory.

A Short Diversion on Node Inflation

- Depends on the foundry
- Technology “node” might reflect other advances (lower leakage or FinFET transistors)
- Not consistent across foundries.

Foundry Node	IDM Node	Min half pitch
7 nm	10 nm	22 nm
5 nm	7 nm	16 nm
3 nm	5 nm	12 nm

Bottom Line: No longer a very meaningful metric

A Short Diversion About ARM Licenses

- **1980s-1990s:** Custom Vector/MPP Market
  - NRE costs not shared by a broader market (hard to recoup dev costs)
  - Technology dev. eclipsed by microprocessor ('killer micro') market
- **1990s-present**: Commodity Microprocessor Market
  - **The Chip is the commodity:** shared by larger desktop/server market
- **ARM play is to make IP the commodity (not the chip)**
  - Share NRE costs with an even larger embedded market
  - Feasible as 64-bit addressing and DP started to appear in embedded
  - Also feasible when clock-rates stopped scaling (arrays of simple cores)
  - Embedded market also enables China, Japan, EU to develop a "domestic technology"
- **Two kinds of licenses**
  - **ISA License:** vendor/country develops microarchitecture, but ISA compliance ensures ALL licensees can rely on common software
  - **IP License:** Can buy a "commodity" IP circuit design from ARMs design library (cost of developing technology is amortized by